УДК 004.5

Модель векторів альтернатив в задачах автоматичного індексування

Ланде Д.В.

Постановка проблеми

В теорії інформаційного пошуку відома так звана векторно-просторова модель представлення документів [1], яка кожному документу ставить у відповідність вектор, так звану «торбу слів» (Bag of the Words). У відповідності із цією моделлю розглядається словник T із всіх термінів, що зустрічаються в документах, що складається із N термінів: $T=\{t_1,t_2,...,t_N\}$. Кожному терміну із документу t_i $(i=1,...,N)$ ставиться у відповідність його вага w_i, величина у діапазоні $[0,1]$. У випадку, якщо термін не входить у деяку множину документів, то його вагове значення приймає нульове значення. У інших випадках, вагу терміна можна розраховувати за відомими алгоритмами. Кожному документу D ставиться у відповідність вектор ваги термінів, що входять до нього: $\bar{W}=(w_1,w_2,...,w_N)$. У відповідності із векторно-просторовою моделлю, для двох документів D_1 і D_2 міра близькості $sim(\bar{W}^1,\bar{W}^2)$ розраховується як косинус кута між відповідними векторами ваги термінів у просторі R^N, тобто, маємо:

$$sim(\bar{W}^1,\bar{W}^2)=\cos(\alpha)=\frac{\sum_{i=1}^{N} w_i^1 w_i^2}{\|\bar{W}^1\|\cdot\|\bar{W}^2\|}.$$

Наведеному підходу притаманні такі недоліки:

1. У випадку роботи з реальними документами необхідно працювати з векторами дуже великої розмірності, що уповільнює вирішення задач, потребує зайвих ресурсів.
2. Необхідність враховувати велику кількість слів, що не несуть змістовного навантаження, при цьому алгоритми розрахунку ваги можуть вносити додаткові помилки.
3. Кут між будь-якими документами при розрахунку міри близькості завжди гострий. (значення косинусу кута між векторами будуть лише невід'ємними), тобто підхід можна застосовувати лише для відносно близьких за тематикою документів.

Мета

Метою роботи є представлення і обґрунтування модифікації векторно-просторової моделі, що базується на представленні документа в просторі альтернатив [2]. Ця модель дозволить позбавитися від зазначених недоліків за рахунок вкладених знань експертів-аналітиків на етапі первинної формалізації. Результати моделювання можна застосовувати при розв'язані задач порівняння і зіставлення окремих тематичних напрямків, зокрема, при оцінці дій уряду і суспільних очікувань, з метою подальшої оцінки синергії суспільства [3], [4].

Обґрунтування

Розглянемо множину альтернатив A, яку будемо розглядати таким чином: $A=\{a_1^+ : a_1^-,\ a_2^+ : a_2^-,...,\ a_N^+ : a_N^-\}$. У цих позначеннях $a_i^+ : a_i^-$ – це граничні значення змістових альтернатив для $i=1,...,N$, наприклад «мир»:«війна», «атака»:«захист», «очищення»:«забруднення» тощо. Цім змістовним, лінгвістичним змінним будемо зіставляти

чисельні вагові значення w_i із діапазону $[-1,1]$. При розгляді окремого документа або тематичного інформаційного потоку значення w_i можуть призначатися, наприклад, виходячи із співвідношення ключових слів у документі, що відповідають крайнім значенням кожної із альтернативи. Множини Наприклад, значенню альтернативі «мир» можуть відповідати ключові слова мир, перемир'я, консенсус, примирення, тощо, а значенню «війна» – такі слова, як «війна», «інцидент», «військова операція», «загострення» тощо. Змістовна задача вибору ключових слів для альтернатив має вирішуватися аналітиками-експертами на етапі «навчання» системи.

Представлення документа у просторі альтернатив аналогічне представленню, наведеному вище, а саме, $\bar{W} = (w_1, w_2, ..., w_N)$.

Близькість окремих як окремих документів, так і цілих тематичних інформаційних потоків розраховується за наведеною вище формулою, але косинус кута може приймати, як позитивні, так і негативні значення. Це обумовлюється розширеним діапазоном можливих вагових значень.

При цьому, маємо такі властивості моделі:

1. Множина альтернатив невелика, тобто векторний простір має обмежену розмірність (практично, не перевищує 100).
2. Всі компоненти кожного вектора не випадкові, їх зміст визначається експертами-аналітиками.
3. Вагові значення елементів вектора визначаються реальним наповненням документів (або тематичних документальних масивів), тому є об'єктивними.
4. Окремі документи, або масиви документів можуть мати протилежне змістовне наповнення, тому і уявний кут між окремими документами (у формулі обчислення близькості) може бути не тільки гострим, а й тупим, що відповідає інтуїтивним уявленням.

Висновки

Наведена модель, що потребує змістовного опрацювання аналітиками-експертами на першому етапі «навчання» системи при виборі альтернатив і відповідних їм ключових слів, дозволяє позбавитися недоліків традиційної векторно-просторової моделі. Представлена модель включає крок навчання системи, яке дозволить у подальшому шляхом мінімальних зусиль з великою достовірністю розрахувати взаємний зв'язок як окремих документів, так і тематичних документальних масивів, зокрема, при розв'язанні задач інформаційної і кібернетичної безпеки, оцінювання суспільних перетворень.

Література

1. Salton G., Wong A., Yang C. S. A vector space model for automatic indexing // Communications of the ACM (1975). – Vol. 18, Issue 11. DOI: https://doi.org/10.1145/361219.361220.

2. Polishchuk V. Technology to Improve the Safety of Choosing Alternatives by Groups of Goals // Journal of Automation and Information Sciences, 2020. – Vol. 20. – P. 2066-76. DOI: 10.1615/JAutomatInfScien.v51.i9.60

3. Zgurovsky M., Lande D., Boldak A. et al. Linguistic Analysis of Internet Media and Social Network Data in the Problems of Social Transformation Assessment. // Cybernetics and Systems Analysis (2021). Volume 57, issue 2. Pages: 228 - 237. DOI: doi.org/10.1007/s10559-021-00348-8.

4. Аналіз сталого розвитку — глобальний і регіональний контексти / Міжнар. рада з науки (ISC) та ін.; наук. кер. проекту М. З. Згуровський. — К. : КПІ ім. Ігоря Сікорського, 2019. — Ч. 1. Глобальний аналіз якості і безпеки життя (2019). — 216 с.